Day One Assignment

Name : Anushka Gupta
SRN : PES2UG20CS060
Class : 5A
Date : 04 July 2022

Problem Statement 1

Check whether the dataset in `gen1.csv` is monotonic and find correlation using the same(spearman/Pearson)

Day One Assignment

Check if the given dataset `gen1.csv` is monotonic or not

In [1]:
```python
# importing Python Packages
import pandas as pd

# reading the dataset
data = pd.read_csv("gen1.csv")

# displaying the dataset
data
```

Out[1]:

	temp	vp	PET	rainfall
0	22.13	23.50	5.87	0.27
1	24.16	22.39	6.52	2.69
2	26.04	24.43	7.21	30.48
3	27.03	36.90	7.26	12.83
4	26.60	45.12	6.88	116.82
...	...	...	...	...
2446	26.95	44.63	5.23	156.34
2447	25.45	37.62	5.28	0.30
2448	26.01	28.74	5.70	NaN
2449	27.12	25.59	6.27	NaN
2450	28.22	28.90	6.64	11.74

2451 rows × 4 columns

In [2]:
```python
data['temp'].is_monotonic
```

Out[2]: False

In [3]:
```python
data['vp'].is_monotonic
```

Out[3]: False

In [4]:
```python
data['PET'].is_monotonic
```

Out[4]: False

In [5]:
```python
data['rainfall'].is_monotonic
```

Out[5]: False

Problem Statement 2

Use the WEKA Explorer and justify the values

1. MCC
2. Kappa Stats
3. ROC Curve Value

For the different pre-defined datasets present under

`C:\\Program Files\\Weka-3-8-6\\data\\diabetes.arff`

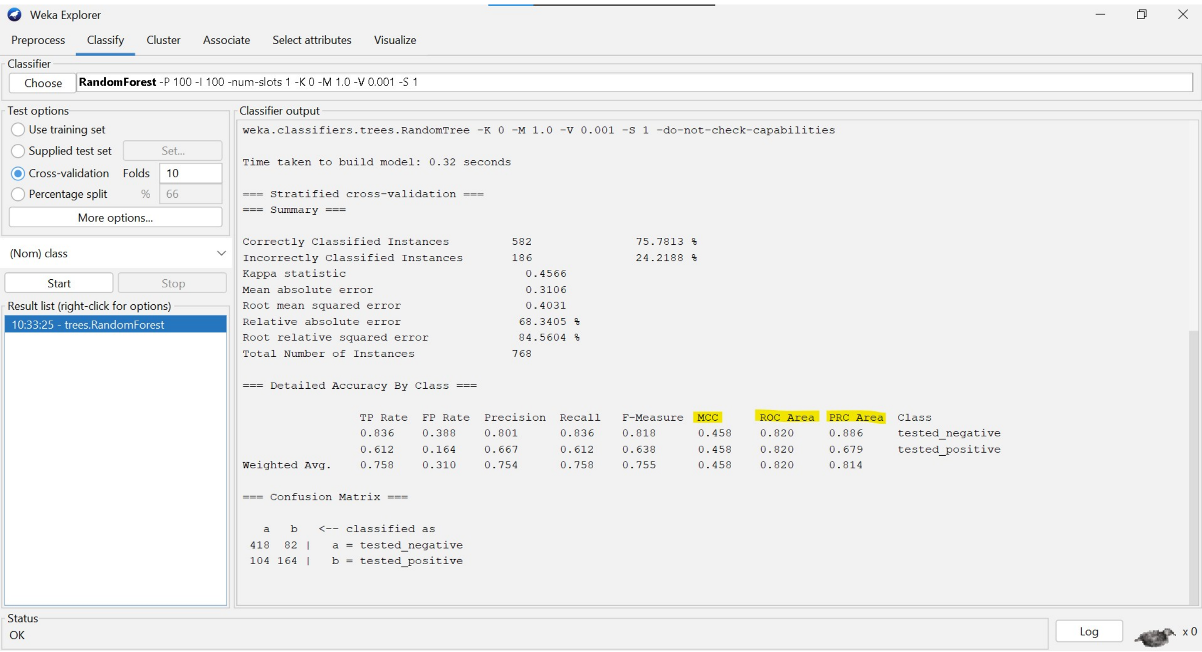
MCC It's **a correlation** between predicted classes and ground truth.

- +1 denotes a perfect model
- -1 denotes a poor model
- 0 denotes that the classifier is no better than a random flip of a fair coin

Kappa Statistics is the ratio of the proportion of times that the appraisers agree (corrected for chance agreement) to the maximum proportion of times that the appraisers could agree (corrected for chance agreement).

ROC Curve Value are frequently used to show in a graphical way the connection/trade-off between clinical sensitivity and specificity for every possible cut-off for a test or a combination of tests. In addition the area under the ROC curve gives an idea about the benefit of using the test(s) in question.

- 0.7 to 0.8 is considered acceptable
- 0.8 to 0.9 is considered excellent
- 0.9+ is considered outstanding



The ROC Curve Value and PRC Area is considered excellent for this data.

MCC value being +0.450 suggests that the model is fairly good.

Problem Statement 3

Calculate Mean, Median and mode for columns rainfall, temp, VP, PET in R

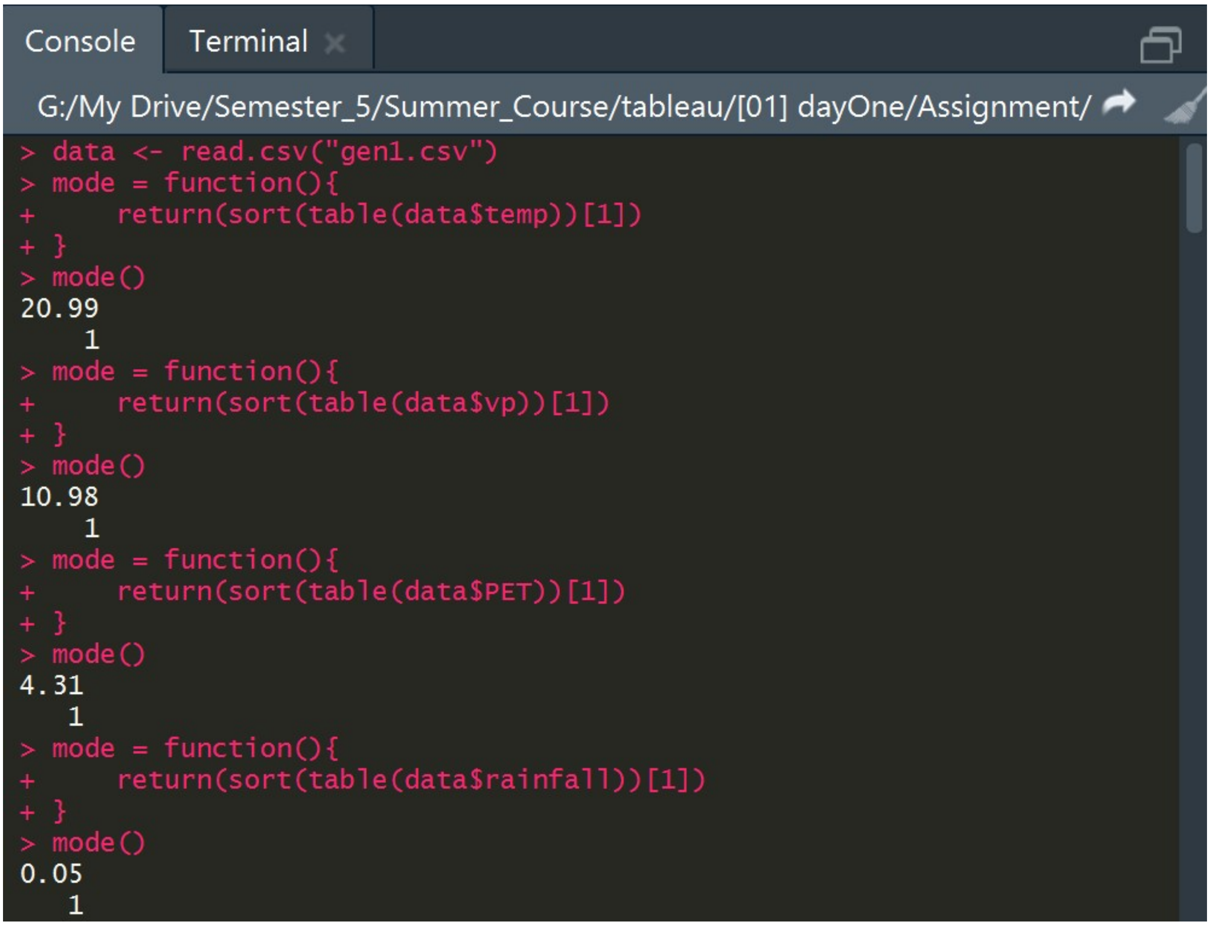
Columns	Mean	Median	Mode
Rainfall	149.5608	78.12	
Temp	25.15173	24.8	

Columns	Mean	Median	Mode
VP	48.51165	46.01	
PET	5.79288	5.46	

```
> data <- read.csv("gen1.csv")
> result.mean <- mean(data$temp)
> print(result.mean)
[1] 25.15173
> result.median <- median(data$temp)
> print(result.median)
[1] 24.8
>
```

```
> result.mean <- mean(data$vp)
> print(result.mean)
[1] 48.51165
> result.median <- median(data$vp)
> print(result.median)
[1] 46.01
>
```

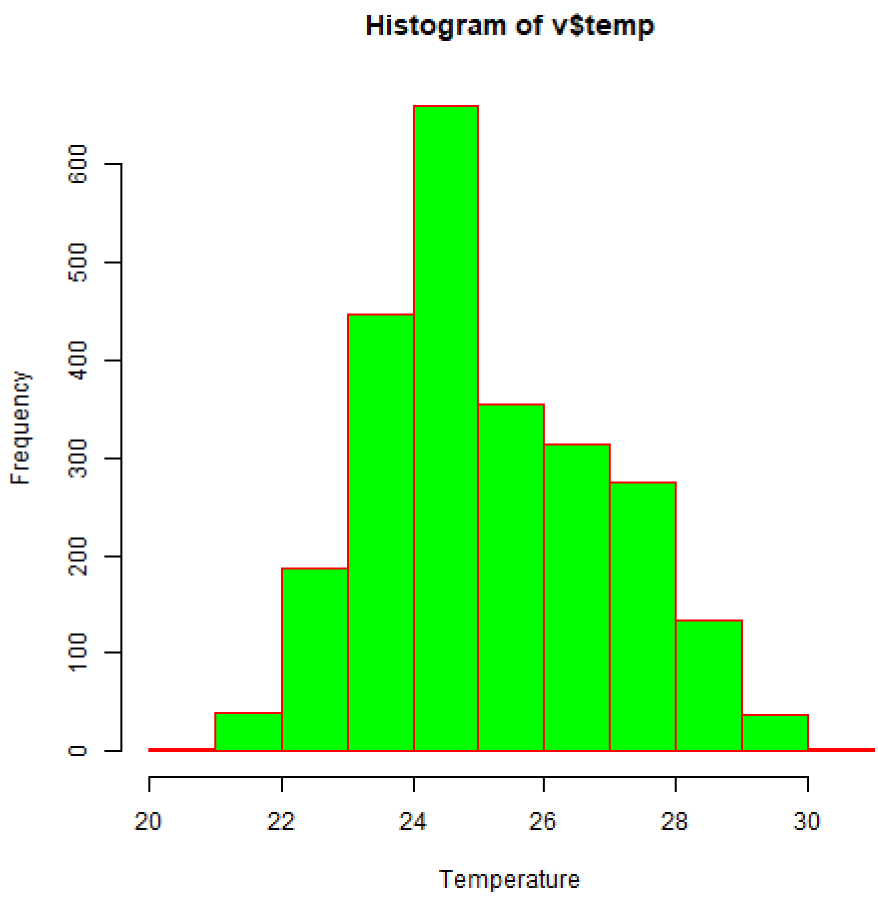
```
> result.mean <- mean(data$rainfall, na.rm = TRUE)
> print(result.mean)
[1] 149.5608
> result.median <- median(data$rainfall, na.rm = TRUE)
> print(result.median)
[1] 78.12
>
```

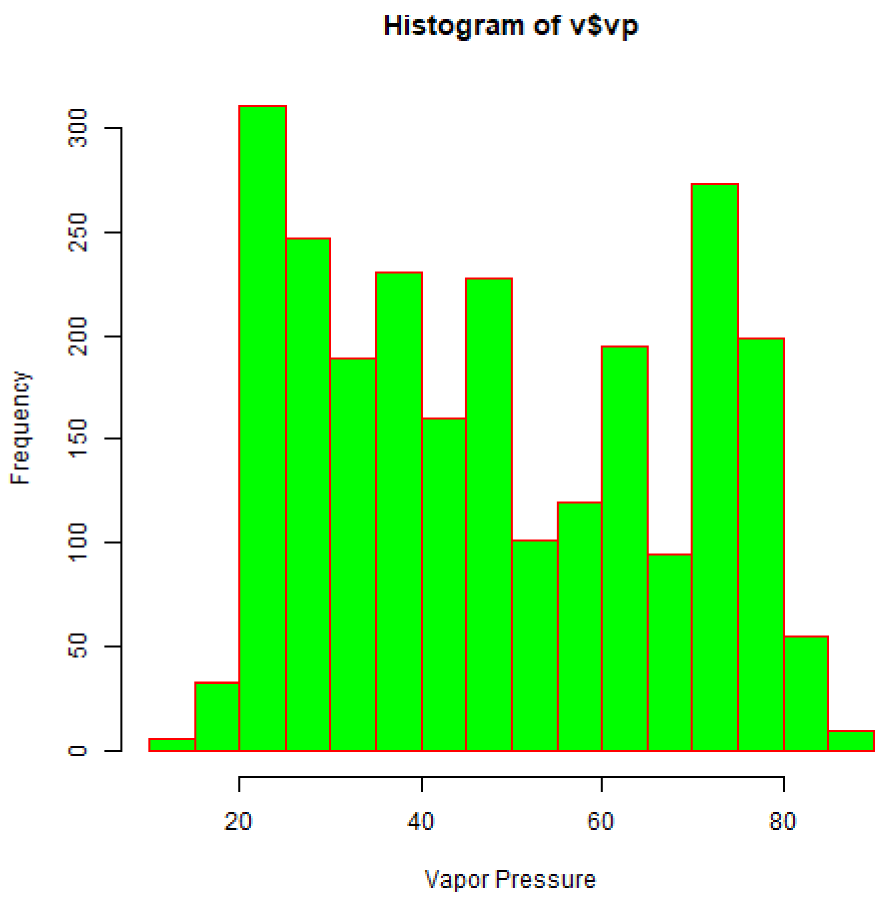


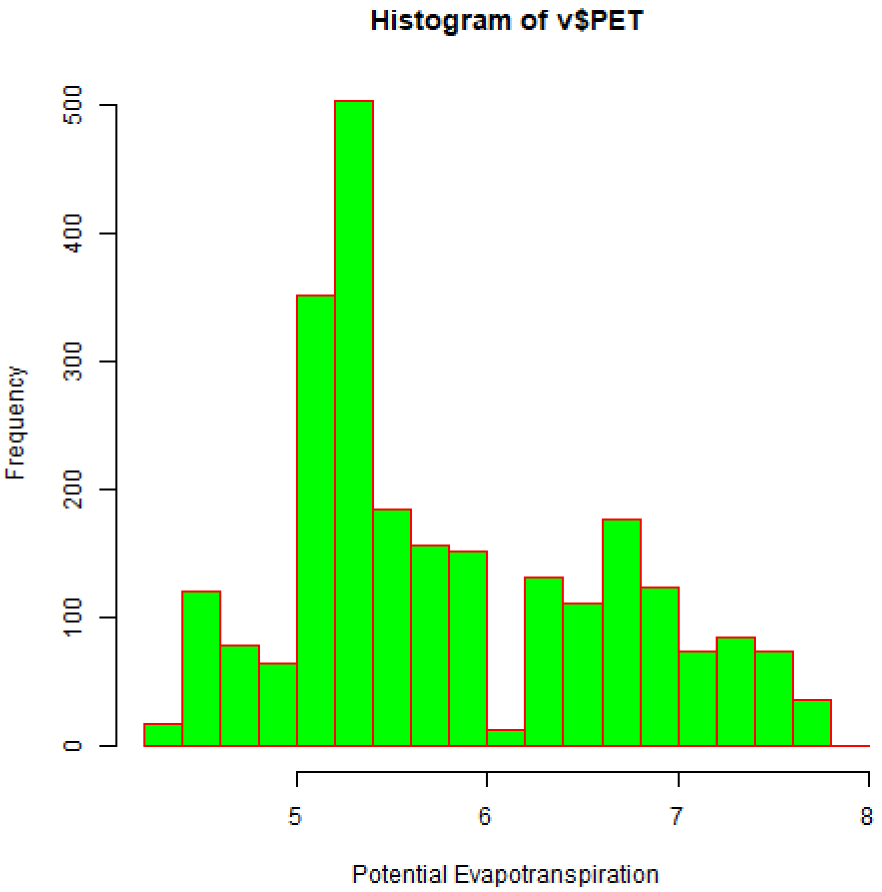
Problem Statement 4

Plot histogram for temp, vp, PET in R

```
> png(file = "temp-histogram.png")
> hist(v$temp, xlab = "Temperature", col = "green",border = "red")
> dev.off()
null device
          1
> png(file = "vp-histogram.png")
> hist(v$vp, xlab = "Vapor Pressure", col = "green",border = "red")
> dev.off()
null device
          1
> png(file = "pet-histogram.png")
> hist(v$PET, xlab = "Potential Evapotranspiration", col = "green",border = "red")
> dev.off()
null device
          1
```


Histogram of v$temp
Frequency
0
100
200
300
400
500
600
20
22
24
26
28
30
Temperature


Histogram of v$vp
Frequency
0
50
100
150
200
250
300
20
40
60
80
Vapor Pressure


Histogram of v$PET
Frequency
0
100
200
300
400
500
5
6
7
8
Potential Evapotranspiration